OAK RIDGE NATIONAL LABORATORY FACT SHEET

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Pushing the Network Simulation Envelope – Extending SSFnet to Terascale Problems

The challenge (and expense) of provisioning high-performance network test beds makes it imperative that their resources be applied in the most cost effective manner possible. Network simulation is a particularly valuable tool for separating and prioritizing worthwhile ideas and tests from experiments that might be a waste of time or resources. The problem has been that simulation of national- or tera-scale test beds has been beyond the abilities of currently available software simulators.

Objectives and Impact

The objective of this project has been to port the JAVA-based SSFNet simulator to the ORNL supercomputing environment, make the resulting package available to researchers throughout the DOE complex, and use it both at ORNL and elsewhere to simulate significant network problems.

Project Components

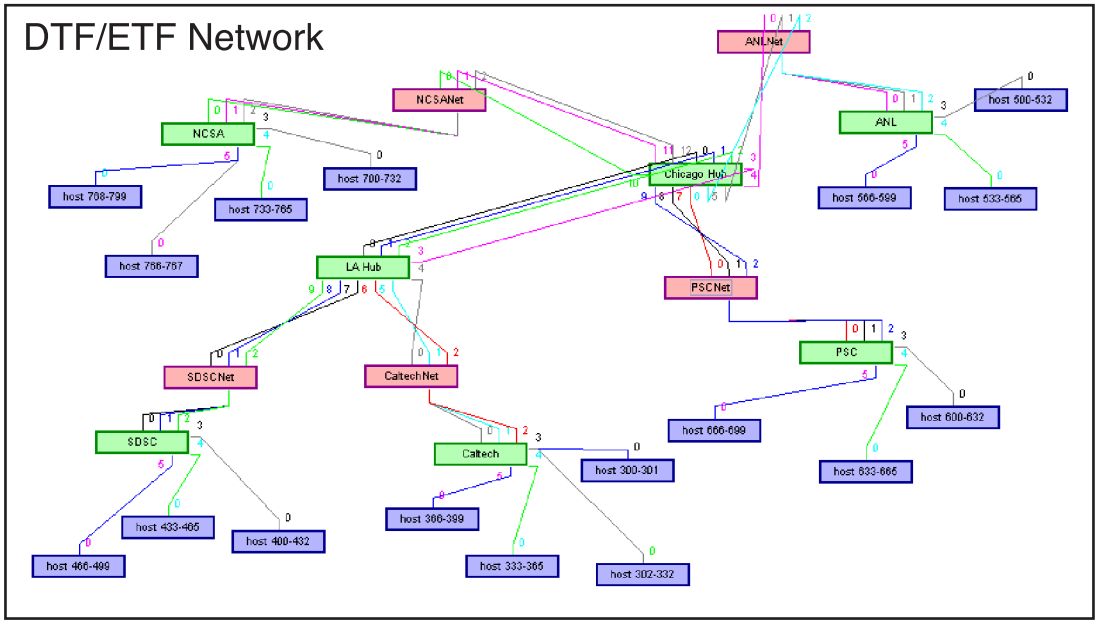
The project is based on two major components. The first is the SSFNet simulation package; the second is a new distributed-memory, multi-threaded kernel, also written in JAVA, which runs in any distributed memory supercomputer with a jdk 1.3 or 1.4-runtime environment. The distributed-memory kernel replaces the shared-memory; proprietary kernel distributed with the SSFNet package and will allow SSFNet to run on almost any multi-node supercomputer or computer cluster. It is based on the C++ distributed-memory kernel written at Dartmouth, and has been rewritten in JAVA.

Although SSFNet is a fairly mature package, it appears in many ways to be more of a research project than a research tool. Thus, we developed two utilities to enable it to model real world networks and interesting problems. The first tool is a visual editor that can convert topologies and network parameters into the thousands of lines of DML (Domain Modeling Language) necessary to specify a network simulation. The second utility is an analog of tcpdump. The current version of SSFNet's tcp stack supports window negotiation, but does not support window scaling, and is limited to 64 kB windows. It was a straightforward hack to force larger windows, but since the built-in tcpdump assumes the same 16-bit window limit, it can't display large-window data. In addition, a bug in the code prevents it from capturing bi-directional flow data. Finally, the extreme dynamic range of the flows in a 10 Gbs WAN made it appropriate to display the data on semi-log scale. Collectively, these encouraged us to develop our own diagnostic data capture and display code.

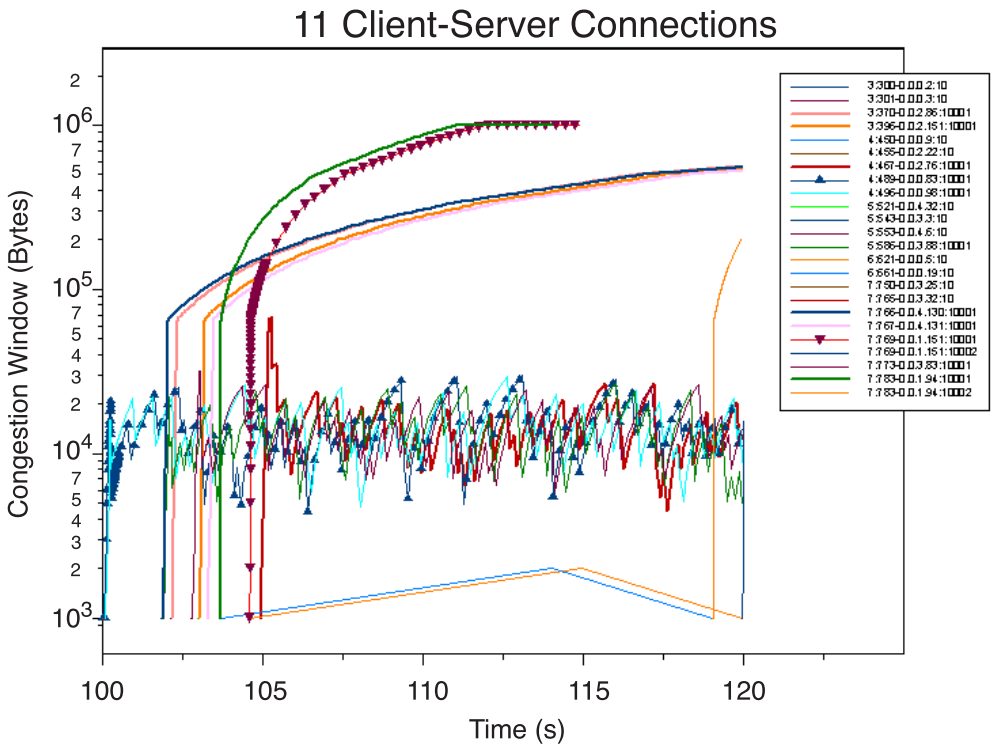
Results

We have started looking for real-world cases to try simulating on SSFNet, and the NSF's DTF/ETF Teragrid network presents an attractive test. The Teragrid (originally called DTF for Distributed Terascale Facility) is the new high-speed research network being funded by NSF, which initially connects clusters of Linux computers at four sites, SDSC, Cal Tech, NCSA, and Argonne National Lab. The network will be extended in the near future to include clusters at NCAR and PSC. The high bandwidth is achieved by using four parallel 10 Gb/s links between the Los Angeles and Chicago hubs, which in turn are linked to each of the computer centers by three parallel 10 Gb/s links. The original DTF architecture consisted of a straightforward mesh between the four sites (N-1 connections at each site). However, the recently funded extension of the network to NCAR and PSC will turn it into a much more general-purpose network.

One feature of the Teragrid, its use of parallel circuits, is important for DOE because it is likely to become the basis for a future network supporting DOE's Ultrascale computing effort, and in general this is a network architecture which is unsupported by general routing protocols.



DTF/ETF network. East-west routers are shown in green, distribution routers in pink, and Linux clusters in blue.



Congestion windows (semi-log scale) for cross-country flows. The uncongested flows are the ones that stay relatively local.